IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Special Issue-9, April 2019 E-ISSN: 2347-2693

Automatic Classification of Research Papers to a Predefined Category Using Machine Learning

Perpetua F Noronha[1*], Prathiba R[2], Gauthami M[3]

[1,2,3] Dept. of Computer Science, Mount Carmel College Autonomous, Bengaluru, India

Corresponding Author: perpetua.noronha @gmail.com, Tel.: 9632600720

DOI: https://doi.org/10.26438/ijcse/v7si9.4751 | Available online at: www.ijcseonline.org

Abstract- With the technology growing exponentially, there are a lot of researches and inventions taking place in all the fields. New innovations and discoveries are put forth in the form of research papers. There are thousands of research papers today that pertain to different disciplines such as Computer Science, Mathematics, Biology, Chemistry etc. Finding papers pertaining to a specific domain is time consuming and a tedious task. Classification of papers to a specific discipline, subject or a category reduces the task of searching. This task if done manually consumes lot of human effort and time where as if done automatically, saves the time of users by preventing them from going through the entire research paper. The proposed work uses a novel strategy to automatically classify the research papers based by analyzing the structure of abstracts of research papers to assign them to a specific predefined discipline. Machine Learning technique is used to build a learning model to learn the properties or characteristics of documents manually, in some cases semi automatically, so that the more it gets trained the more efficient will be the model to predict or classify the test documents. Support Vector Machine (SVM) algorithm is used to vectorize the training data set and plot them in an n-dimensional space and then to find the hyper plane that will separate the data into a predefined category. The data is then learnt and later used to categorize the data. The performance of SVM is compared with Naïve Bayes and Decision Tree algorithms also. The experimental result proves the outstanding performance of SVM to predict the category of research papers over the other two algorithms mentioned above. The main objective of the proposed work is to develop a system that has the ability to learn from a training set of data, improvise from the experiences without explicitly programming for it and later classify any research paper given to it into a discipline.

Keywords- Support Vector Machine (SVM), Bag-Of-Words (BOW), Machine Learning (ML)

I. INTRODUCTION

Large volume of digital data is being constantly generated at unprecedented and exponential rate as mentioned in the work [1]. This Voluminous data which is termed as Big Data is collected and studied in numerous domains, from engineering sciences to social networks, commerce, bimolecular research, and security [2]. In this age of Big Data, textual data is speedily growing and is available in many different languages and has high social and economic significant value. Text and data mining techniques and text analytics is needed to utilize this potential. Text mining and text analytics gained importance for leveraging the information from unstructured data. Business has always wanted to derive insights from information in order to make better, smarter, real time, fact-based decisions. Through techniques such as categorization, entity extraction, sentiment analysis and others, text mining extracts the useful information and knowledge hidden in text content. With an iterative approach, an organization can successfully use text analytics to gain insight into content-specific values such as sentiment, emotion, intensity and relevance. There is a need to manage various types of documents more effectively as text or document classification. Automated text classification has been considered as a vital method to manage and process a vast amount of documents in digital forms that are widespread and continuously increasing. In general, text classification plays an important role in information extraction and summarization, information retrieval (IR) and natural language processing fields and is a process of assigning the predefined categories to text documents. It is this demand for depth of knowledge that has fuelled the growth of text classification tools and platforms. Natural Language processing is an area of computer science concerned with the interactions between computers and natural languages, in particular how to program computers to process and analyze large amounts of Natural Language data is defined in the work [3]. It enables a computer to read and analyze textual information. Natural Language Processing interprets the meaning of the text and identifies, extracts, synthesizes and analyses relevant facts and relationships that directly answer your questions. Sophisticated Natural Language Processing (NLP) algorithms, allow recognizing similar concepts – even if they’ve been expressed in very different ways, or with different spellings.

Machine Learning is one of the prime driving factors of the Big Data revolution. It has the potential to provide data driven insights, decisions, and predictions using past experience or sample data expressed in the paper [4].

Numerous frameworks have been developed to work alongside with Machine Learning algorithms on large text sets and conquer the challenges faced by Machine Learning with Big Data. Machine Learning approach can be used to classify text. It allows us to measure the results of a classification algorithm in an objective way. Once the unstructured text is converted into semi-structured or structured data using Text Mining or Text analytics, Machine Learning techniques/algorithms can be applied to get new unknown informative and meaningful data discussed in the work [5]. A trainable model can be obtained by the application of a trainable Machine Learning algorithm in the collection of documents. The two broad categories of Machine Learning are Supervised and Unsupervised Machine Learning. Supervised is the learning under supervision and find their applications in processing and analyzing variety of data. The most important characteristic of supervised learning is their ability in annotating training data as given by the author [6]. Unsupervised learning models a set of inputs: labeled examples are not available. Unsupervised Machine Learning is more closely aligned with what some call as true Artificial intelligence. It is the idea that a computer can learn to identify complex processes and patterns without a human to provide guidance along the way.

II. RELATED WORK

Text classification has been an important research area since 1950. In this era of internet there is a tremendous rise in the availability of textual data from various digital media. One of the important tasks in information retrieval and Natural Language processing is automatic classification of text i.e a process of assigning the predefined categories to text documents.

In the research paper [9], the authors have proposed a hybrid model which uses k-near neighbor and support vector machine algorithms. The paper presents a two phase approach to test on the datasets. In the first phase, the k-near neighbor model is applied to calculate the neighbor list which is training phase. The k-near neighbor finds the distance between every centroid which is used in second phase. The second phase Support Vector Machine classifier uses the above input to classify into the respective categories. The proposed model works well on large datasets and produced good results. The authors in the work [10], implement a different feature extraction method using support vector machine classifier. In this approach, the time domain and frequency domain are analyzed by translating and scaling co-efficient using Discrete Wavelet Transform (DWT). After that the data is given to a statistical model as Multi-Dimensional Scaling (MDS) to find out the common words and classified into the respective class.

Authors in the work [11] propose a classification algorithm based on Logical Analysis of Data (LAD). Data is encoded into binary form, this is done by using the training set for extracting values for each field, that split each field into binary attributes. The particular binary attributes constitute a support set, and are combined for generating logical rules called patterns. Patterns are used to classify each unclassified record, on the origin of the sign of a weighted sum of the patterns activated by that record. The results of the proposed work produced good results and less time for small datasets.

In the research paper [12], author presents a novel fuzzy support vector machine (FSVM) tool or a variant of FSVM called modified fuzzy support vector machine (MFSVM). This variant is to classify the credit approval problem. In FSVM, each sample is given a fuzzy membership which denotes the attitude of corresponding point toward one class. The membership function which is a hyperbolic tangent kernel grips the impreciseness in training samples. In MFSVM, the victory of the classification lies in proper selection of the fuzzy membership function which is a function of center and radius of each class in feature space and is represented with kernel. The kernel used in MFSVM is hyperbolic tangent kernel. This kernel allows lower computational cost and higher rate of optimistic eigen values of kernel matrix which eases several limitations of other kernels. In research paper [13], authors propose an innovative and active approach to guess the Bayesian probability. The proposed classifier, called EnBay, emphases on choosing the minimal set of long and not overlapped patterns that best complies with a conditional-independence model, based on an entropy-based evaluator. Additionally, the probability approximation is distinctly tailored to each group. This model works on two periods partition the attribute set into a minimal number of large subsets so that their conditional dependence, given an arbitrary class, is minimized, then choose frequent item sets considered by conditionally independent attribute sets.

In the research work [14], authors describe a novel and cost-sensitive Naive Bayes approach by drawing an inference between the true classification’s possibility of assigning to the class of interest and the cost-sensitive threshold. This method concludes that the classification is based on the inferred order relation. The authors in [15] propose a combined approach of k-nearest neighbor classification method and support vector machine classification algorithm which uses Bayesian vectorization to convert documents into numeric form. Then in training phase training data points are plotted into vector space of the SVM. Then support vectors of each class are recognized and the residual training data points are discarded. Now in classification stage new unlabeled data point in charted into the same vector space of support vectors which gained from the training phase. Then estimate average distance for each categories by using Euclidean distance formula and then after define the group of new unlabeled data point based on the shortest average distance between the support vector of the category and the new data point.

The authors in the paper [16] presents an integrated approach of combining K-Nearest Neighbor(KNN) text classification algorithm and document based centroid dimensionality reduction (CentroidDR) method in R to derive a classifier model which proves more effective than existing ones. In the work [17] authors present the use of WordNet tool with the existing KNN classifier to derive a precise and efficient classification model.

The research work in [18] presents a Bayesian classification approach combined with SVM to arrange the text content automatically by using class specific components. The classifier gives promising results but at the cost of time.

III. METHODOLOGY

High efficiency in the performance of document classification is expected in the proposed model. The same model is been proposed in [7] which uses multinomial Naïve Bayes technique because of its efficiency and popularity. The proposed work [8] illustrates the best performance of multinomial naïve bayes technique than Knearest neighbor. The proposed model compares SVM, Naïve Bayes and Decision Tree, Machine Learning techniques for classification and a very simple and novel strategy for text transformation in order to achieve better accurate text classification results as compared to the results of [7] work. The proposed classification model is developed in the following steps as given in the Figure 1. The life cycles of the proposed model undergoes three phases training, testing and validate. All the labeled and unlabeled documents are preprocessed first. In the training phase, a model is built and trained using labeled data. In the test phase the model is tested by using the unlabeled class data whose labels are known but not used. Finally in the validate phase, the model is tested with an entirely new set of prepared data whose class label is unknown.

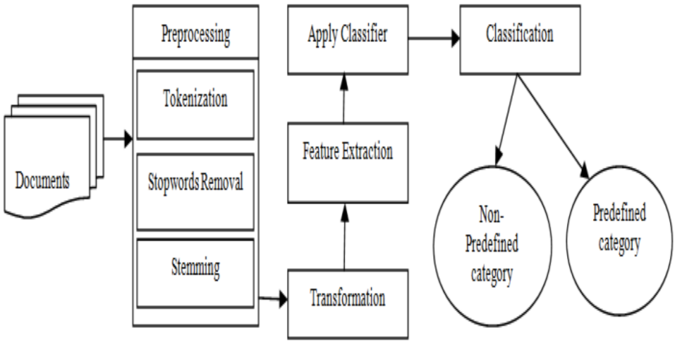

Figure 1. Classification Model Architecture

III. 1 PRE-PROCESSING

Preprocessing is the process of converting raw data into a clear format. In this proposed work preprocessing is applied on all input documents in all the phases to present the text in a clear format. The basic steps performed in this stage are:

III. 1.1 TOKENIZATION

Tokenization is the act of breaking up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. Tokens can be individual words, phrases or even whole sentences. In the process of tokenization, some characters like punctuation marks are discarded. The tokens become the input for another process like parsing and text mining.

III. 1.2 STOPWORDS REMOVAL

Stop words are the English words which do not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence. For example, the words like the, he, have etc. Such words are already captured this in corpus named corpus. Text may contain stop words like ‘the’, ‘is’, ‘are’. Stop words can be filtered from the text to be processed. There is no universal list of stop words in NLP research; however the NLTK module contains a list of stop words.

III. 1.3 STEMMING

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Stemming is important in natural language understanding (NLU) and natural language Research Paper Classifier Page 18 processing (NLP). Stemming is a part of linguistic studies in morphology and artificial intelligence (AI) information retrieval and extraction. Stemming and AI knowledge extract meaningful information from vast sources like big data or the Internet since additional forms of a word related to a subject may need to be searched to get the best results.

III. 2 TRANSFORMATION

Transformation is document representation in which the given plain text document is converted into an instance with fixed number of attributes. From among the various methods available, Bag-Of-Words (BOW) is the most popular method used. In the proposed model strings belonging to Computer Science field are stored manually as a vocabulary list. The model uses term frequency weighting method to determine the number of time a word exists in the document and then a weigh matrix is created. The strings in the document are compared with vocabulary list and the frequency of their occurrences in the document and also the percentage of their occurrences in the document is computed. Finally a vector space model is created which represents each document as a vector with n number of selected features.

III. 3 FEATURE EXTRACTION

Feature extraction is the process of creating entirely new data set which is a subset of original dataset. In the proposed model feature extraction is performed by using random function to divide the data set into trained and test data. The trained data will have all the independent variables and also the dependent class label. The test data will have only the independent variables. Factorization then used to get the numeric representation for the class label. In the proposed model the class label is represented as 0 if the document belongs to the predefined category else it is 1 if the document belongs to some other category.

III. 4 TEXT MINING METHODS

Number of text mining methods in data mining have been proposed and used such as: Classification, Clustering, Information retrieval, Topic discovery, Summarization, Topic extraction. Text classification is implemented in the proposed model using three different Machine Learning algorithms like SVM, Naïve Bayes and Decision Tree.

III. 5 TRAINING PHASE

The input to this phase is the preprocessed text document and the output is a trained model which can be used to test and classify new set of documents. The training phase of the proposed model is depicted in the Figure 2.

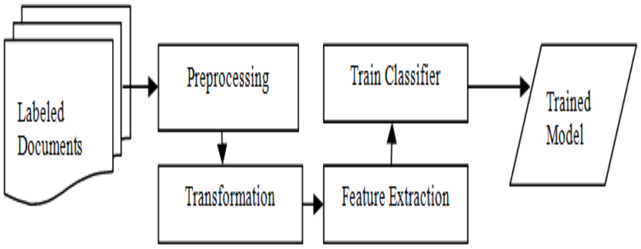

Figure 2. Training Stage

III. 6 TESTING PHASE

In this stage the trained model is tested for performance and evaluated as shown in Figure 3. The inputs to this phase are unlabeled documents prepared during preprocessing stage and their associated class labels used for validation stage. In the evaluation process the input is the predicted class labels of the testing documents from the classification step and the actual associated class labels.

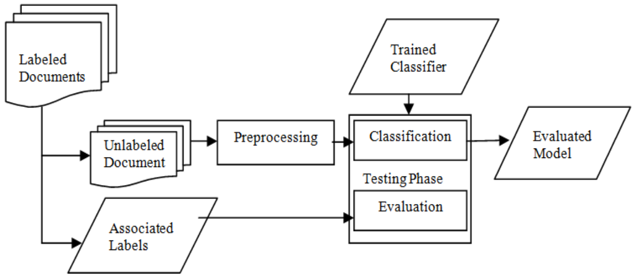

Figure 3. Testing Phase

III. 7 VALIDATE PHASE

In this stage the trained, tested and evaluated model is used to classify any given new data with unknown class labels as shown in Figure 4. This phase assigns any new unlabelled document to a predefined category.

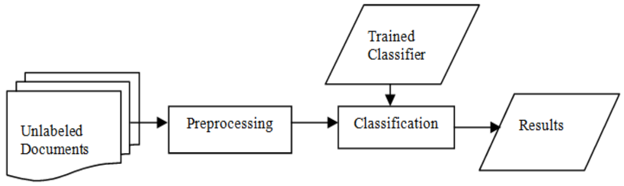

Figure 4. Validate Phase

IV. EXPERIMENTAL RESULTS

The experiment was carried out using the following system setup given in the Table 1.

Table 1. System Setup

Operating System	Windows7
CPU	Intel Core i5 6th Generation processor
RAM	8 GB
Language	Python

The work is carried out by considering around 125 abstracts of research papers belonging to Computer Science and other categories. Attributes like abstract ID, total number of words, frequency of occurrences of those words in the document which had a match in the predefined category dictionary, the percentage of frequency and the category to which the document belongs or not are computed for each abstract and stored in excel sheet. Using the feature extraction method the whole dataset is split into 80% of training dataset and 20% of test dataset. Results of implementing all the three Machine Learning algorithms SVM, Naïve Bayes and Decision Tree is evaluated in terms of Precision, Recall, F-Score and Accuracy given the Table 2.

Table 2. Performance evaluation of all the three algorithms

Results	**SVM**	**Naïve Bayes**	**Decision Tree**
Accuracy	0.95	0.92	0.95
Precision	0.96	0.93	0.96
Recall	0.96	0.93	0.96
F-Score	0.96	0.93	0.96

Support Vector Machine is declared as the best classifier for the proposed work because it can be used to distinguish inseparable linear data in a more efficient way than the other two algorithms.

V. CONCLUSION AND FUTURE SCOPE

The proposed work in this paper presents a classification model to classify the given abstracts into a predefined category or not. The experiment results confirm the best performance of SVM classifier method over other Naïve Bayes and Decision Tree methods. The proposed model can be improved by using feature selection, vectorization methods for document representation and then implementing machine learning algorithms like SVM to classify research papers into multiple categories. The proposed work can be enhanced to classify the research documents into multiple categories by considering methodologies and experimental results along with the abstract of the paper.

REFERENCES

[1] Junfei Qiu, Qihui Wu, Guoru Ding , Yuhua Xu and Shuo Feng, “*A survey of Machine Learning for big data processing*”, Qiu et al. EURASIP Journal on Advances in Signal Processing (2016) 2016:67, DOI 10.1186/s13634-016-0355-x.

[2] A Sandryhaila, JMF Moura, “*Big data analysis with signal processing on graphs: representation and processing of massive data sets with irregular structure*”, IEEE Signal Proc Mag 31(5), 80–90 (2014).

[3] Diksha Khurana, Aditya Koli, Kiran Khatter, Sukhdev Singh, “*Natural Language Processing: State of The Art, Current Trends and Challenges*”, arxiv.org/pdf/1708.05148, August 17, 2017.

[4] D. Saidulu, Dr. R. Sasikala, “Machine *Learning and Statistical Approaches for Big Data: Issues, Challenges and Research Directions*”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 21 (2017) pp. 11691-11699.

[5] Sheetal Shimpikar, Sharvari Govilkar, “*A Survey of Text Summarization Techniques for Indian Regional Languages*”, International Journal of Computer Applications (0975 – 8887), Volume 165 – No.11, May 2017.

[6] Vladimir Nasteski, “*An overview of the supervised machine learning methods*”, Journal of advances in information technology, DOI10.20544/HORIZONS.B.04.1.17.P05,UDC 04.85.021:519.718, Jan 4, 2017.

[7] Mowafy M, Rezk A, El-bakry HM, “*An Efficient Classification Model for Unstructured Text Document*”. American Journal of Computer Science and Information Technology, Vol.6 No.1: 16, ISSN 2349-3917, 2018.

[8] Rajeswari RP, Juliet K, Aradhana, “*Text Classification for Student Data Set using Naive Bayes Classifier and KNN Classifier*”, International Journal of Computer Trends and Technology, – Volume 43 Number 1 – January 2017.

[9] M. Kepa, J. Szymanski, “*Two stage SVM and k-near neighbor text documents classifier*”, In the Proceedings of the 6th International Conference, on Pattern Recognition and Machine Intelligence PReMI 2015, Warsaw, Poland, June 30-July 3, 2015, DOI: 10.1007/978-3-319-19941-2_27, pp.279-289.

[10] R. C. Barik and B. Naik, **"***A Novel Extraction and Classification Technique for Machine Learning using Time Series and Statistical Approach***",** International Journal of Innovative Research in Computer and Communication Engineering, DOI: 10.15680/IJIRCCE.2018.0605115, Vol. 6, Issue 5, May 2018.

[11] R. Bruni and G. Bianchi, "*Effective Classification Using a Small Training Set Based on Discretization and Statistical Analysis*"**,** IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. X, XXXXX 2015.

[12] A. Chaudhuri, "*Modified fuzzy support vector machine for credit approval classification*," Journal AI Communications, Volume 27 Issue 2, April 2014, Pages 189-211.

[13] E. Baralis, L. Cagliero, and P. Garza, "*EnBay: A novel pattern-based Bayesian classifier*", IEEE Transactions on Knowledge & Data Engineering, pp. 2780-2795, vol. 25, Dec. 2013.

[14] X. Fang, "*Inference-Based Naive Bayes: Turning Naive Bayes Cost-Sensitive*", IEEE Transactions on Knowledge and Data Engineering 25(10):2302-2313 · October 2013.

[15] C. H. Wan, L. H. Lee, R. Rajkumar, and D. Isa, "*A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine*", International Journal of Innovative Research in Science, Engineering and Technology, Volume 3, Special Issue 3, March 2014.

[16] Maganti Syamala, Dr N J Nalini, Lakshamanaphaneendra, Dr. R Ragupathy, “*Comparative Analysis of Document level Text Classification Algorithms using R*”, IOP Conf. Series: Materials Science and Engineering 225 (2017) 012076 doi:10.1088/1757-899X/225/1/012076.

[17] M. Parchami, B. Akhtar, and M. Dezfoulian, "*Persian text classification based on K-NN using wordnet*”, book Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Preface (pp.283-291), DOI: 10.1007/978-3-642-31087-4_30.

[18] Autade Sushma G., Dr.Gayatri M.Bhandari, “*Text Categorization based on SVM and Bayesian Classification Approach Using Class-Specific Features*”, International Journal of Advanced Research in Computer Engineering & Technology,Volume 06, Issue 06, June 2017, ISSN: 2278 – 1323.